

私立東海大學資訊工程與科學研究所


碩士論文

指導教授：林祝興 博士

Dr. Chu-Hsing Lin

針對二元文件影像保護之竄改還原技術

A Tamper Recovery Scheme for the Protection of  
Binary Document Images

The seal of the National Central University Library is visible in the background, featuring the university's name in both Chinese and English.

研究生：鄭掄元

(Lun-Yuan Cheng)

中華民國九十七年六月

東海大學碩士學位論文考試審定書

東海大學資訊工程與科學系 研究所

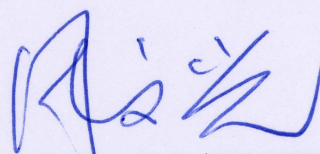
研究生 鄭 掄 元 所提之論文

針對二元文件影像保護之竄改還原技術

經本委員會審查，符合碩士學位論文標準。

學位考試委員會

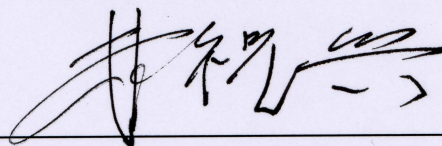
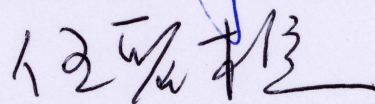
召 集 人



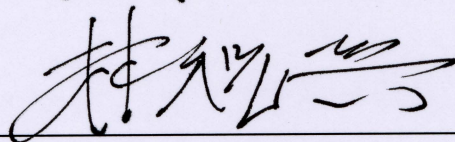
簽章

委

員



指 導 教 授



簽章

中華民國 97 年 6 月 30 日

# 致 謝

本篇論文之所以能順利完成，要感謝的人實在很多，首先要感謝我的指導老師 林祝興教授，兩年來在學業上悉心指導，獲益良多，使我在學術研究的領域中有所成長。另外，感謝 劉榮春老師在論文與實驗上指導與啟發，本文乃得以順利完成，僅向兩位恩師致上學生最高的敬意與謝意，也感謝口試委員們撥冗指正論文中錯誤與提供許多寶貴的意見，使本篇論文能更加完整。

在學期間也感謝實驗室學長鎮宇、佳穎、三隻熊、菊人、仁傑、coji、丸子學姐與郁文關於研究和課業上給予很多建議與幫助；實驗室同學猴子、newtype、DC與懋樺彼此為我加油打氣與安慰；學弟妹們建廷、美君、志捷與宗哲在我沮喪時適時給予鼓勵與歡笑，謝謝你們。

研究生生涯使我成長許多，家人的鼓勵與支援是我最大的精神支柱，僅將此論文獻給我最敬愛的父親、母親、哥哥，與所有關心我的親友們。今日，終於完成碩士學位。在此，致上我最大的感謝與祝福予所有的師長、家人、朋友與學弟妹們，願目前本研究能對未來研究者有所裨益。

鄭掄元 謹上 2008/7

# Abstract

Nowadays ever-increasing amount of data are transmitted through the Internet, and they are in danger of tampering. So, protection of the integrity of the data becomes very important. Until present, few methods have been published that successfully address important issues on file verification, and detection of document tampering. In the thesis, we propose a new method to protect and recover binary document images from tampering. Black skeletons are calculated first and then information is embedded accordingly in the host file. The embedded information can be extracted from the stego file anytime later to compare with the original black skeletons for file verification. Afterward, we can recover the tampered area by using our proposed method. The experimental results show that our proposed method is effective in detecting locations of tampering and is capable of recovering binary document images.

**Keywords and phrases:** Binary document images, File integrity, Black skeleton, Tamper detection,

# 摘要

現今有越來越多的資料需要透過網路來互相傳遞，而這些傳輸中的資料是含有被竄改的潛在危險。因此，保護數據的完整性是非常重要的。目前，已經有了一些方法成功地解決檔案確認和偵測文件竄改的問題。在這篇論文中，我們提出了一種新的方法，利用計算黑色骨架的和嵌入資訊在原始的文件中，來保護和回復遭受竄改的二元文件影像。其中可以從原始的文件中取出黑色骨骼來做為嵌入的資訊。並採用我們所提出的方法進行破壞部份的復原工作，實驗結果表明，我們所提出的方法是可以有效的在偵測出遭受竄改的位置，並且回復的二元文件影像。

**關鍵詞：**二元文件影像、檔案完整性、黑色骨架、竄改偵測

# Contents

<b>Contents .....</b>	<b>3</b>
<b>Figures.....</b>	<b>5</b>
<b>Tables .....</b>	<b>8</b>
<b>Chapter 1 Introduction.....</b>	<b>9</b>
<b>Chapter 2 Backgrounds.....</b>	<b>11</b>
<b>2.1. Data Hiding.....</b>	<b>11</b>
<b>2.2. The Classification of Data Hiding.....</b>	<b>12</b>
<b>2.3. Binary Document Image.....</b>	<b>14</b>
<b>2.4. Data Hiding on Document Image .....</b>	<b>14</b>
<b>Chapter 3 The Proposed Method .....</b>	<b>16</b>
<b>3.1. Skeleton Extraction .....</b>	<b>16</b>
<b>3.2. Embedding Process .....</b>	<b>17</b>
<b>3.3. Detection Process.....</b>	<b>19</b>
<b>3.4. Recovery Process .....</b>	<b>20</b>

<b>Chapter 4 Experimental Result .....</b>	<b>22</b>
<b>4.1. Experimental Environment.....</b>	<b>22</b>
<b>4.2. Measurement Tools (PSNR).....</b>	<b>22</b>
<b>4.3. Results With Different Stego Document Images .....</b>	<b>23</b>
<b>4.4. Analysis of Different Stego Document Image .....</b>	<b>28</b>
<b>4.5. Effect of Tampering Percentage .....</b>	<b>29</b>
<b>4.6. Analysis of Tampering Percentage Effect.....</b>	<b>32</b>
<b>4.7. Embedding Catacity.....</b>	<b>33</b>
<b>Chapter 5 Conclusions and Future Works.....</b>	<b>35</b>
<b>Bibliography .....</b>	<b>36</b>

# Figures

<b>Figure 1 The Classification of Data Hiding.....</b>	<b>12</b>
<b>Figure 2 Black Skeleton.....</b>	<b>17</b>
<b>Figure 3 Embedding Process.....</b>	<b>18</b>
<b>Figure 4 Bad Blocks to Embed Information .....</b>	<b>18</b>
<b>Figure 5 The Odd/Even Value .....</b>	<b>19</b>
<b>Figure 6 Stego Document Be Tampered .....</b>	<b>20</b>
<b>Figure 7 Detect The Tampering Part.....</b>	<b>20</b>
<b>Figure 8 Recovery Process .....</b>	<b>21</b>
<b>Figure 9 Attack and Recovery of A Typed English Document.....</b>	<b>25</b>
<b>9(a) Host Document.....</b>	<b>23</b>
<b>9(b) The Black Skeletons .....</b>	<b>24</b>
<b>9(c) Stego Document.....</b>	<b>24</b>
<b>9(d) The Word “Document” Is Shown to Be Cropped .....</b>	<b>24</b>
<b>9(e) The Recovered English Document .....</b>	<b>24</b>



<b>Figure 10 Attack and Recovery of A Typed Chinese Document.....</b>	<b>27</b>
<b>10(a) Host Document.....</b>	<b>25</b>
<b>10(b) The Black Skeletons .....</b>	<b>25</b>
<b>10(c) Stego Document.....</b>	<b>26</b>
<b>10(d) The Word “現代” Is Shown to Be Cropped.....</b>	<b>26</b>
<b>10(e) The Recovered Chinese Document.....</b>	<b>26</b>
<b>Figure 11 Attack and Recovery of A Handwritten Document .....</b>	<b>28</b>
<b>11(a) Host Document.....</b>	<b>27</b>
<b>11(b) The Black Skeletons .....</b>	<b>27</b>
<b>11(c) Stego Document.....</b>	<b>27</b>
<b>11(d) The Word “interest” Is Shown to Be Cropped.....</b>	<b>28</b>
<b>11(e) The Recovered Handwritten Document .....</b>	<b>28</b>
<b>Figure 12 Attack and Recovery of The Tampering Handwritten Document.....</b>	<b>31</b>
<b>12(a) Cropping 25%.....</b>	<b>30</b>
<b>12(b) Recovery 25% .....</b>	<b>30</b>

<b>12(c) Cropping 50%</b> .....	<b>30</b>
<b>12(d) Recovery 50%</b> .....	<b>31</b>
<b>12(e) Cropping 75%</b> .....	<b>31</b>
<b>12(f) Recovery 75%</b> .....	<b>31</b>
<b>Figure 13 Embedding Capacity</b> .....	<b>33</b>
<b>Figure 14 Embedding Capacity Be Cropped 75%</b> .....	<b>34</b>
<b>Figure 15 Embedding Capacity Less Than The Sizw of Black Skeleton</b> .....	<b>34</b>
<b>Figure 16 Incorrect Black Skeleton</b> .....	<b>34</b>

# Tables

**Table 1 The Analysis of PSNR (Different Stego Document Image) .....29**

**Table 2 The Analysis of PSNR (Different Percentage Cropping Attack).....32**

# Chapter 1

## Introduction

Nowadays people are used to transmitting documents via the Internet. Digital documents are pervasively used in our daily life and the security issue of them has become more and more important in current society [1, 2]. Presently, cryptographic technology is booming and it provides useful tools for document integrity and authentication services.

Digital documents are vulnerable to illegal modifications by unauthorized parties when transmitted through the Internet. In order to protect the important information of original document images, information data are embedded in the digital documents. If tampered, the legitimate users can recover the documents by using the embedded information.

Besides, it is more difficult to protect text documents than color or gray images. Since most text documents are in binary format [3-5] and thus there might be not enough capacity to secretly embed information into the text documents. It is very challenging to undetectably embed information into a text document and extract the information to recover the document at a later time. A few related researches in this area can be found in the literature [6-8].

In this thesis, we proposed a novel method to detect tampering of binary documents based on skeletons [9]. We compute the black skeletons of the document and embed the information of skeletons in it. The proposed scheme can be used to

recover the binary documents.

This thesis is organized as follows. In Chapter 2, background of the thesis is given. In Chapter 3, we describe our proposed method. The experimental results are shown in Chapter 4. Chapter 5 concludes this thesis.

# Chapter 2

## Backgrounds

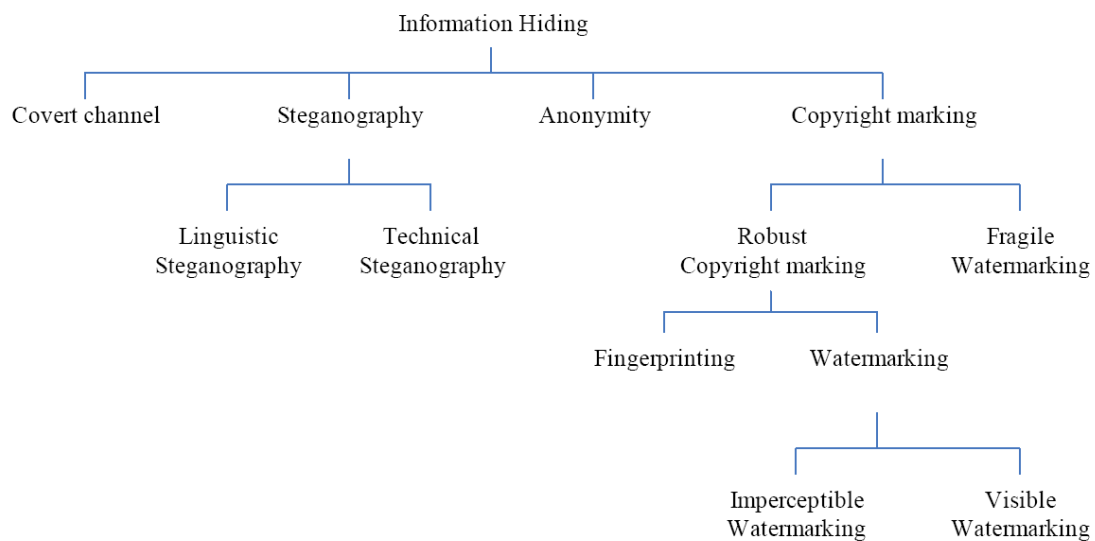
### 2.1 Data Hiding

Digital watermark and steganography are classified in the fields of data hiding [10]. Data hiding and cryptosystem are different. Cryptosystem is message encrypted computing. In fact, the encrypted text will let people become alert, and more prone to trying to destroy the text. Without the decryption key, the information will not be unfolded. Therefore, we can use cryptography to protect text document. And it is like adding a strong lock. Data Hiding is to embed information in the media. And if outside people do not know there are other hidden messages in the media, the purpose of secret communications is achieved. The technology of data hiding can be applied in digital media to protect the intellectual property rights and enhance the integrity of digital media.

Steganography is composed by "steganos" and "graphein". The secret correspondence has a very long history in the evolution of steganography, The most common form of steganography is invisible ink and microdot in the spy activities. At old time, the invisible ink often uses milk, fruit juice or urine to write on the thesis. The secret message will emerge by applying heat on the thesis. Microdot is quite popular during the Second World War. Intelligence officers often make text image smaller and hide it in an unimportant letter to convey secret messages.

## 2.2 The Classification of Data Hiding

The first session of the International Information Hiding Conference [11] in 1996 defined the topic about information hidden technology. Petitcolas et al. [12] classified information hiding by different applications into four branches shown in Fig. 1.



**Figure 1.** The classification of data hiding

Following is the applications of information hiding technology:

- 1. Ownership Assertion:** Embed information in the media to identify the owner of ownership.
- 2. Integrity Verification:** Check the media itself for completion and correctness, the main purpose is to verify whether the media was attacked or not.
- 3. Content Labeling:** A function of digital subtitles describes or indexes briefly on the content of digital products. It is usually applied to digital video animation.
- 4. Usage Control:** To control the number of access for specific user on the digital products. If access to reach the number, the data will be broken and failure occurs,

and users will not be able to access the digital content.

**5. Convert Communication:** It is one of the most ancient of the application for hiding secret information in order to get a security communication.

**6. Content Protection:** it is mainly used to protect the contents of the original data and avoid an attacker to tamper the original data.

Information hiding technology on digital image needs to meet the following conditions:

**1. Imperceptibility:** Imperceptibility means the embedded information will not degrade the quality of the original image. In other words, the embedded information must be visually indistinguishable from the original image.

**2. Undetectable:** Stego image is not easy to be analyzed to reveal the hidden information by computing and statistical methods.

**3. Undeletable:** The hidden information is non-removable and it should be hard to detected and modified. In addition, unhidden information will not only destroy the aesthetic sense of images but also expose the location of the embedded information.

**4. Robustness:** Robustness focuses on the resistance to attempts of removing information. A robust scheme can extract enough information from attacked images to reconstruct the embedded data. Image attacks include geometric attacks (such as rotation, scaling, transpose, etc.) and non-geometric attacks (such as Gaussian noise, sharpening, and contrast adjustment, etc.). But robustness is usually inversely proportional with the image quality.

**5. Capacity:** Capacity is the largest amount of information hidden in original image. Generally, if we hide larger information, it is easier to detect.

**6. Unambiguous:** When verifier got stego image, he must be able to extract the hiding information effectively.



## **2.3 Binary Document Image**

A binary image is a digital image that has only two possible values for each pixel. Typically the two colors used for a binary image are black and white, though any two colors can be used. For the gray and color images, many researchers have used digital watermark to protect property rights and integrity and have published many related protection schemes. Binary images have less capacity to hide information than gray and color images, which can effectively use rich colors to achieve the purpose of information hiding. In recent studies, some scholars propose the embedded schemes on binary images to find the best embedding locations to embed information that will not significantly reduce the quality of the original binary images.

## **2.4 Data Hiding on Document Image**

Steganography or information hiding [9] can be used to encrypt important information into host documents to protect them. Steganography is different from cryptography that obscures information and makes it unreadable without help of particular knowledge. Information hiding secretly embeds undetectable information into media. The advantage of steganography over cryptography is that embedded information does not attract any attention. However, if the amount of embedded secret information is too much, it will not only slow down the transmission speed but also catch the eyes of attackers, who would try to tamper the document and obtain the secret information.

As a result, when the amount of secret information is less, the stego document is more secret. In the thesis, we use some characteristics of original document images for recovery. We extract black skeletons from the host image. In text document, the part of black is usually less than the part of white. Therefore, we can encrypt less amount of secret information in the extracted black skeleton that constitutes just small part of the binary document image.

# Chapter 3

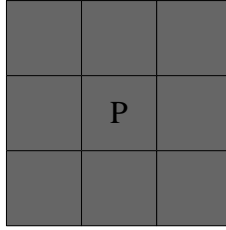
## The Proposed Method

In this chapter, we introduce the proposed binary document recovery scheme, which is based on the skeleton extraction technique. We compute the black skeletons of a host document image by using the skeleton extraction technique, and select the locations which can be used to embed information, and then embed the information into the host document image. The embedded information can be extracted later on to recover the tampered binary documents. To embed and extract the information, we use the method proposed by Lu et al. in [4].

### 3.1. Skeleton Extraction

The skeleton extraction procedure is described here. First, we have to extract black skeletons from the host image. In our method, the candidate locations to embed information are the whole block of image excepting the skeleton parts.

The black skeletons are located as follows. At first we divide the image into overlapping  $3 \times 3$  blocks, from left to right and from up to down of the image. If pixels values of any  $3 \times 3$  block are all 0 (or black) as shown in Fig. 2, the center pixel value of this block is 0 and the pixel is called a black skeleton. Finally, all black skeletons are combined to form one skeleton. In this way, the skeleton is successfully extracted.



**Figure 2.** Black skeleton.

P is a black skeleton since pixels values of the  $3 \times 3$  block are all 0.

## 3.2. Embedding Process

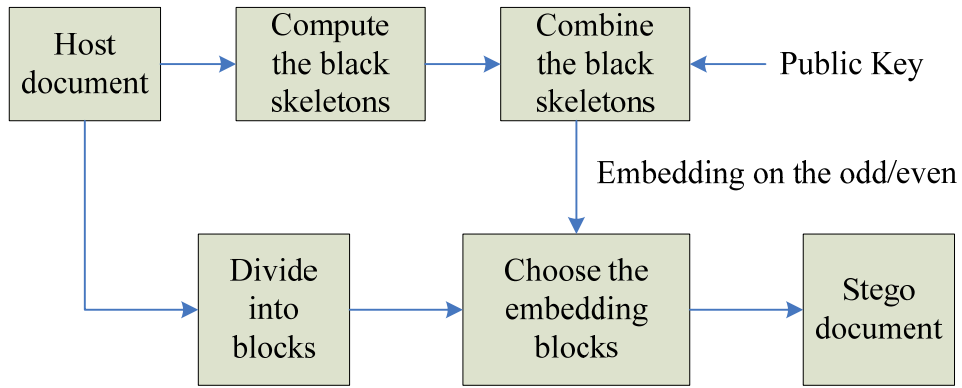
In this section, we describe the embedding process of the proposed scheme (see Fig. 3). By using the above skeleton extraction technique on the host document, we obtain black skeletons.

**Input:** Host document, and black skeletons.

**Output:** The stego document.

The embedding step is described as follows.

- E1.** Use skeleton extraction algorithm described in 3.1 to compute black skeletons of the host document.
- E2.** Combine the black skeletons to form a single skeleton  $IE$ .
- E3.** Divide the host document into blocks, count pixels inside blocks and record odd/even count of the block pixels.
- E4.** Choose the best embedding locations to embed the information  $I_E$ . The embedding steps are performed from left to right and from up to down to produce a new document image, which is the stego document.

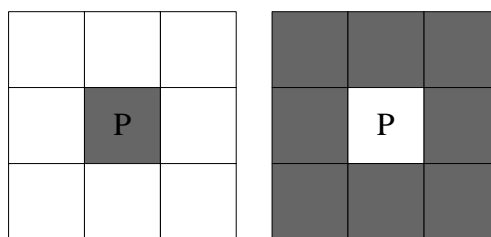


**Figure 3.** Embedding process

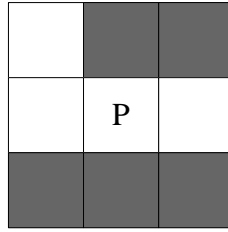
The information embedding process on the odd/even block is described here:

If the number of black pixel is odd, 1 is embedded. And if the number of black pixel is even, 0 is embedded.

Process of choosing embedding locations is explained as follows. Fig. 4 shows blocks consisting of all all-black or all-white value, except the center part. Blocks like these would not be used, since any change of a pixel value in these blocks is visually detectable. Fig. 5 shows a good location to embed information. The “1” is embedded in this block since it is an odd block.



**Figure 4.** Bad blocks to embed information.



**Figure 5.** The odd-even value

Since the odd-even value is odd, we embed “1” in this block.

### 3.3. Detection Process

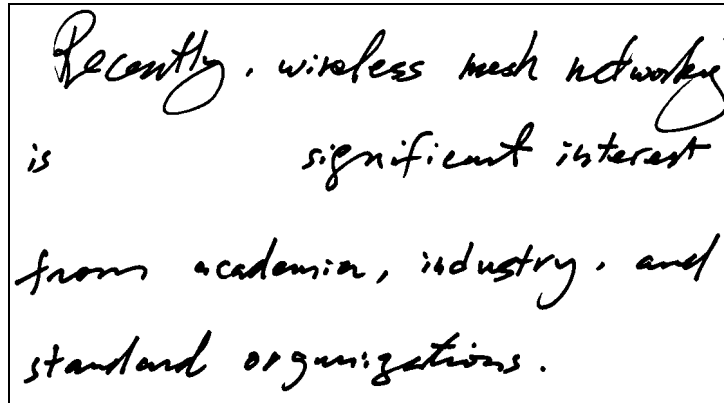
In this section, we describe the document image detecting process of our proposed scheme.

**Input:** The stego document.

**Output:** The detected stego document.

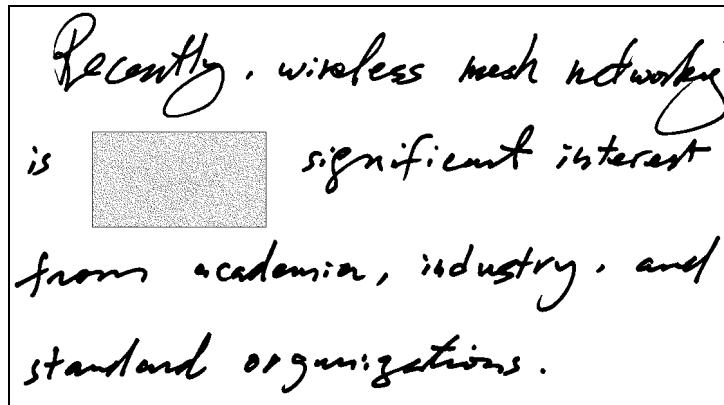
The detection step is described as follows.

- D1.** We divide the stego document image into overlapping  $3 \times 3$  blocks to calculate the embedding locations. We extract embedding information from stego document image by odd/even value.
- D2.** We do skeleton extracting process on stego document image. Then we compare the skeleton information from stego document image with the re-calculation skeleton information.
- D3.** If the two skeletons are the same, it means that the stego document is not been attacked. If different frames of skeleton information are found, the stego document has been attacked (see Fig. 6).
- D4.** We compare pixel values of two skeleton data to find the tampering part (see Fig. 7) and perform image recovery process by the original skeleton information.



Recently, wireless mesh networking  
is significant interest  
from academia, industry, and  
standard organizations.

**Figure 6.** Stego document be tampered



Recently, wireless mesh networking  
is [redacted] significant interest  
from academia, industry, and  
standard organizations.

**Figure 7.** Detect the tampering part

### **3.4. Recovery Process**

In this section, we describe the document image recovery process of our proposed scheme (Fig. 8).

**Input:** The tampered stego document.

**Output:** The recovered document.

The recovery process is as follows.

**R1.** Divide the tampered stego document into blocks and calculate the embedding

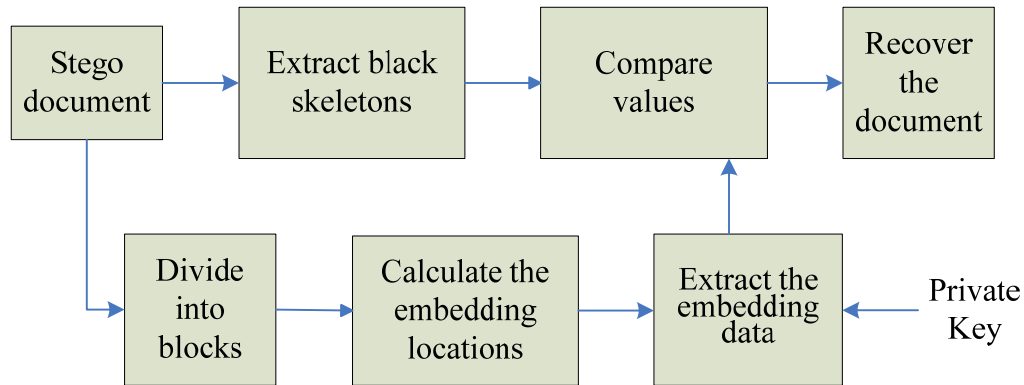
locations.

**R2.** Extract the black skeleton.

**R3.** Compare values of the black skeleton and the tampered stego document.

Locations of tampering of document are detected if black skeletons are found at places where pixels values of the tampered document are white.

**R4.** Replace locations of tampering with values of the black skeletons to recover the binary document image.



**Figure 8.** Recovery process



# Chapter 4

## Experimental Results

### 4.1. Experimental Environment

The experiments were conducted in Windows XP operation system. We used JBuilderX IDE, Java Advanced Imaging (JAI) package library. The programming language is JAVA jdk 1.6.0. Photoshop 8.0 was used for image tampering. Handwritten documents and typewritten English documents are used as host documents.

### 4.2. Measurement Tools (PSNR)

To measure the quality of the stego image, we compute its Peak Signal to Noise Ratios (PSNR) value. Stego images with higher PSNR are more similar to the original images. The PSNR is defined as below:

$$PSNR = 10 \times \log_{10} \frac{255^2}{MSE} dB$$

Where MSE is the mean square error of the two images:

$$MSE = \left( \frac{1}{m^2} \right) \sum_{i=0}^{m-1} \sum_{j=0}^{m-1} (\alpha_{ij} - \beta_{ij})^2$$

Where  $\alpha_{ij}$  s stand for pixels of original image,  $\beta_{ij}$  s stand for pixels of stego image.

### 4.3. Results with Different Stego Document Images

Fig. 9(a) shows an English host document image. Fig. 9(b) shows the extracted black skeleton. Fig. 9(c) is the stego document. In Fig. 9(d), the word “Document” of the stego document is shown to be cropped. As shown in Fig. 9(e), after the image recovery process, the stego document is successfully recovered.

Fig. 10(a) shows a Chinese host document image. Fig. 10(b) shows the extracted black skeleton. Fig. 10(c) is the stego document. In Fig. 10(d), the word “現代” of the stego document is shown to be cropped. As shown in Fig. 10(e), after the image recovery process, the stego document is successfully recovered.

Fig. 11(a) shows a handwritten host document image. Fig. 11(b) shows the extracted black skeleton. Fig. 11(c) is the stego document. In Fig. 11(d) the word “interest” of the stego document is shown to be cropped. As shown in Fig. 11(e), after the image recovery process, the stego document is successfully recovered.



(a)

**Binary Document Images Authentication**  
*Chu-Hsing Lin, Wen-Kui Chang, Yu-Yi...*  
*Department of Computer Science and I...*  
*Tunghai University, 407 Tai...*  
*{chlin, wkc, g95280074, g95280...*

(b)

**Binary Document Images Authentication**  
*Chu-Hsing Lin, Wen-Kui Chang, Yu-Yi...*  
*Department of Computer Science and I...*  
*Tunghai University, 407 Tai...*  
*{chlin, wkc, g95280074, g95280...*

(c)

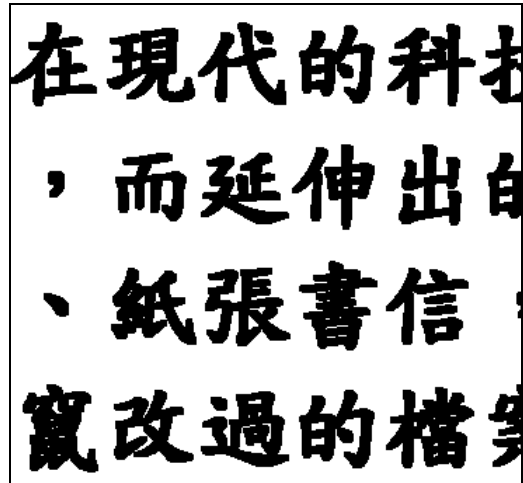
**Binary Document Images Authentication**  
*Chu-Hsing Lin, Wen-Kui Chang, Yu-Yi...*  
*Department of Computer Science and I...*  
*Tunghai University, 407 Tai...*  
*{chlin, wkc, g95280074, g95280...*

(d)

**Binary Document Images Authentication**  
*Chu-Hsing Lin, Wen-Kui Chang, Yu-Yi...*  
*Department of Computer Science and I...*  
*Tunghai University, 407 Tai...*  
*{chlin, wkc, g95280074, g95280...*

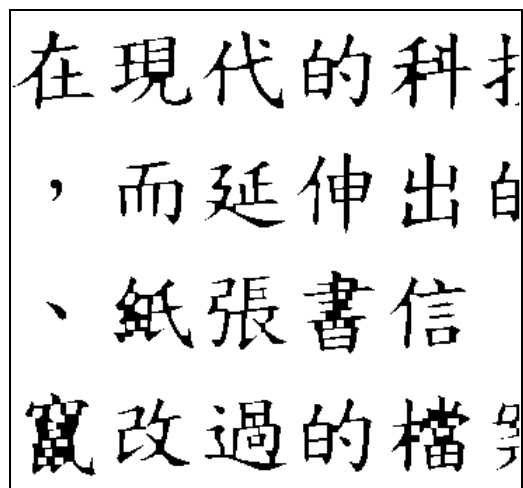
(e)

**Figure 9.** Attack and recovery of a typed English document of 911×315 pixels. (a) host document, (b) the black skeletons, (c) stego document, (d) the word “Document” is shown to be cropped, (e) the recovered English document.



在現代的科技  
，而延伸出自  
、紙張書信  
竄改過的檔案

(a)



在現代的科技  
，而延伸出自  
、紙張書信  
竄改過的檔案

(b)

在現代的科技  
，而延伸出自  
、紙張書信  
竄改過的檔案

(c)

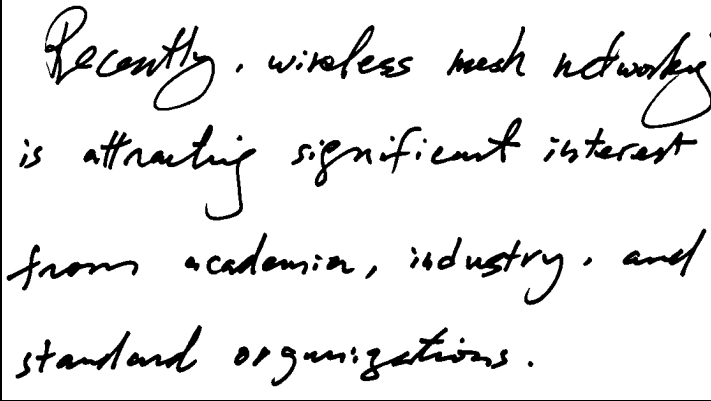
在  的科技  
，而延伸出自  
、紙張書信  
竄改過的檔案

(d)

在現代的科技  
，而延伸出自  
、紙張書信  
竄改過的檔案

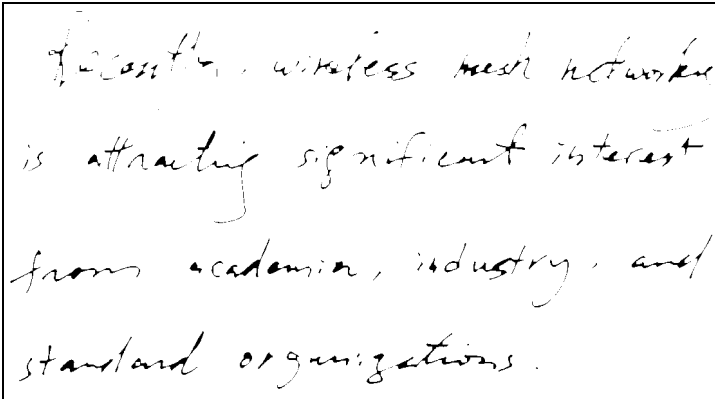
(e)

**Figure 10.** Attack and recovery of a typed Chinese document of 364×344 pixels. (a) host document, (b) the black skeletons, (c) stego document, (d) the word “現代” is shown to be cropped, (e) the recovered Chinese document.



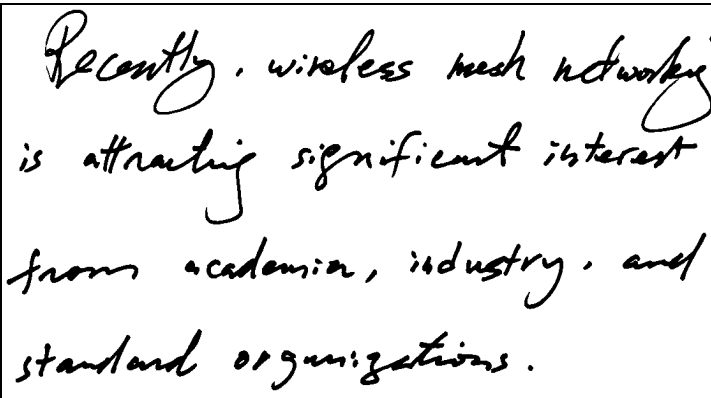
Recently, wireless mesh networking is attracting significant interest from academia, industry, and standard organizations.

(a)



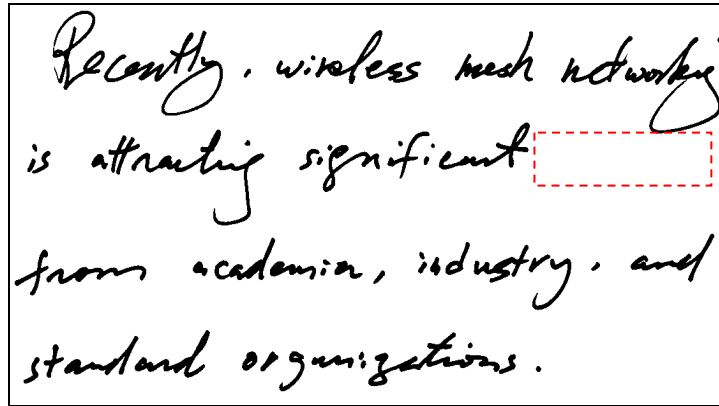
Recently, wireless mesh networking is attracting significant interest from academia, industry, and standard organizations.

(b)



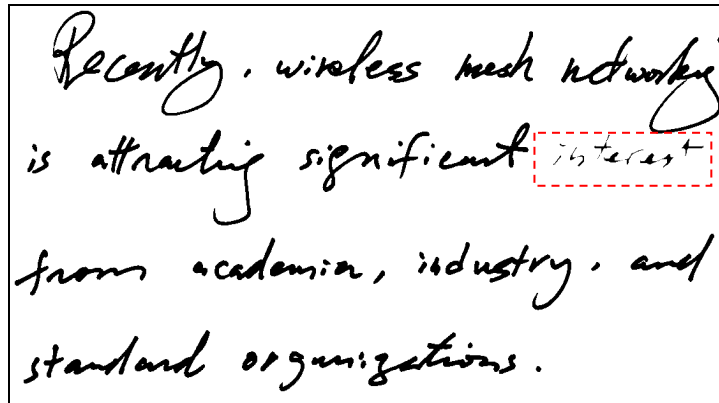
Recently, wireless mesh networking is attracting significant interest from academia, industry, and standard organizations.

(c)



Recently, wireless mesh networking  
is attracting significant  
from academia, industry, and  
standard organizations.

(d)



Recently, wireless mesh networking  
is attracting significant  
from academia, industry, and  
standard organizations.


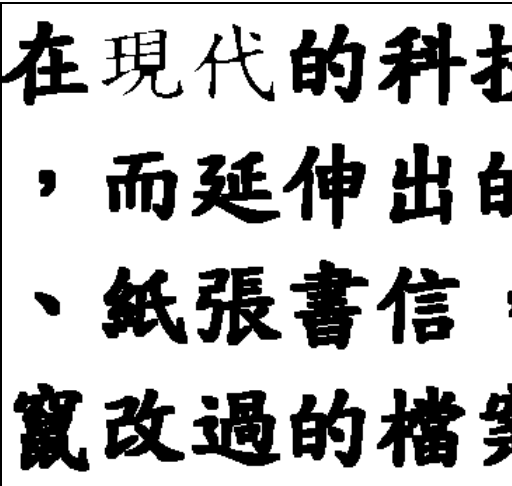
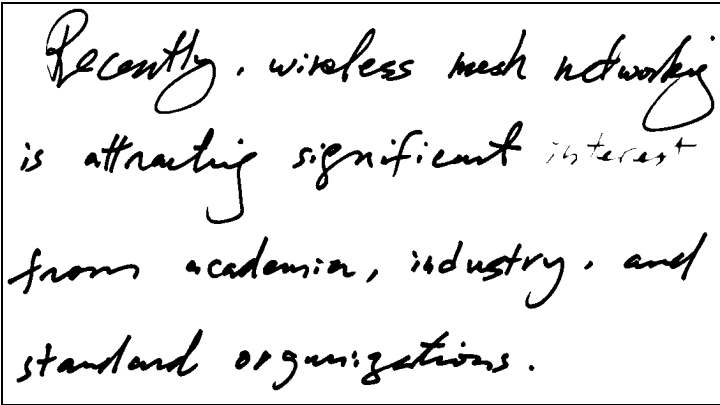
(e)

**Figure 11.** Attack and recovery of a handwritten document of 1125×624 pixels. (a) host document, (b) the black skeletons, (c) stego document, (d) the word “interest” is shown to be cropped, (e) the recovered handwritten document.

#### 4.4. Analysis of Different Stego Document Image

Table 1 lists the recovered stego document images and PSNR values of the English and Chinese document images and the handwritten document image.

**Table 1.** The analysis of PSNR (different stego document image)

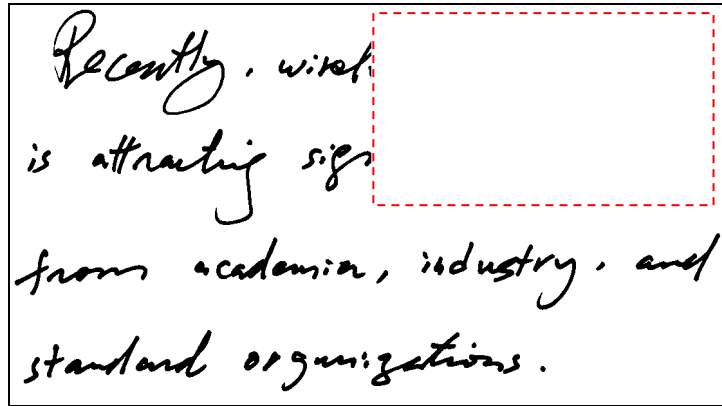
Recovery document image		PSNR
English document image	 <p><b>Binary Document Images Authentication</b> <i>Chu-Hsing Lin, Wen-Kui Chang, Yu-Yi...</i> <i>Department of Computer Science and I...</i> <i>Tunghai University, 407 Tai...</i> <i>{chlin, wkc, g95280074, g95280...</i></p>	21.548
Chinese document image	 <p>在現代的科技 ，而延伸出的 、紙張書信 竄改過的檔案</p>	17.999
Handwritten document image	 <p>Recently, wireless mesh networking is attracting significant interest from academia, industry, and standard organizations.</p>	24.969

## 4.5. Effect of Tampering Percentage

In addition, we do experiment of the recovery rate. We used handwritten document images as the experimental objects. After 25% 50% and 75% cropping

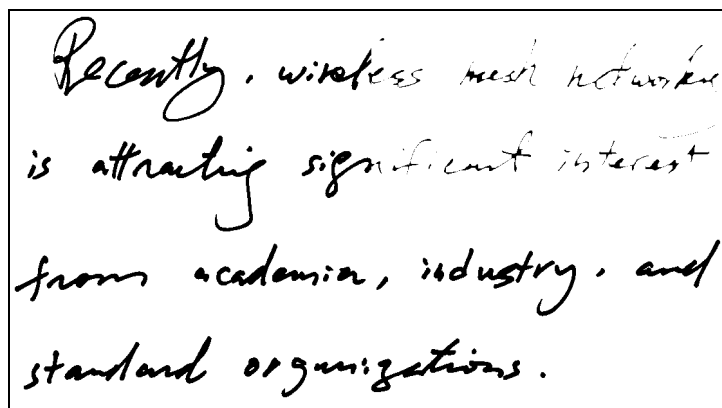


attacks on the image, we effectively recover the binary documents by our proposed scheme.



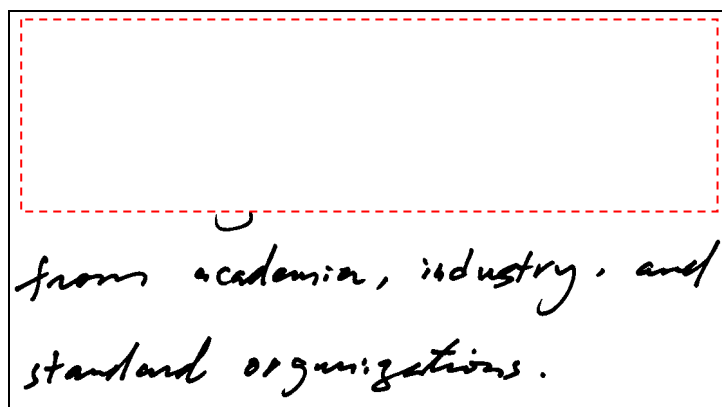
Recently, wireless  
is attracting sig  
from academia, industry, and  
standard organizations.

(a)



Recently, wireless mesh networks  
is attracting significant interest  
from academia, industry, and  
standard organizations.

(b)



from academia, industry, and  
standard organizations.

(c)

Recently, wireless mesh networks  
is attracting significant interest  
from academia, industry, and  
standard organizations.

(d)

from academia,  
standard organiz

(e)

Recently, wireless mesh networks  
is attracting significant interest  
from academia, industry, and  
standard organizations.

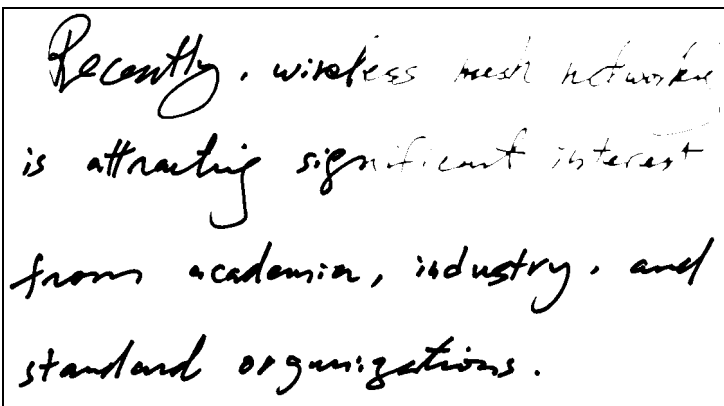
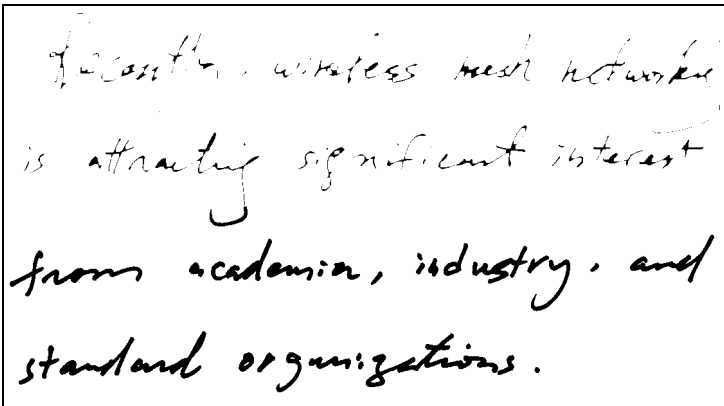
(f)

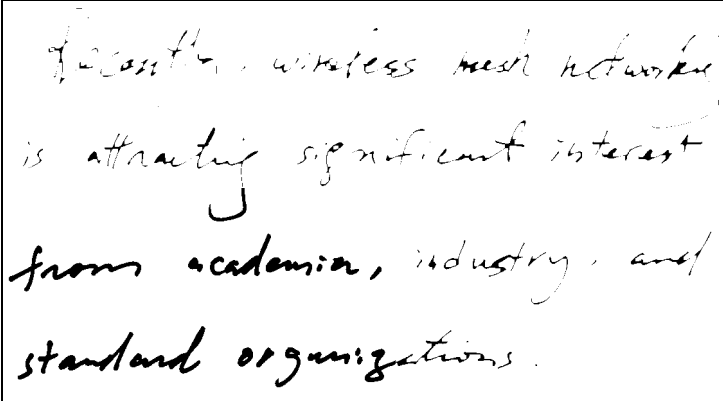
**Figure 12.** Attack and recovery of the tampering handwritten document of 1125×624 pixels (a) cropping 25%, (b) recovery 25%, (c) cropping 50%, (d) recovery 50%, (e) cropping 75%, (f) recovery 75%

## 4.6. Analysis of Tampering Percentage Effect

Table 2 lists the recovered stego document images which are recovered from 25%, 50% and 75% cropping attacks. Their PSNR values are shown in the last column.

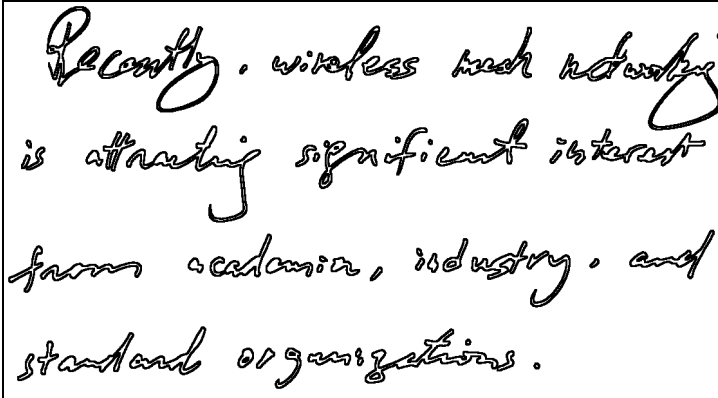
**Table 2.** The analysis of PSNR (different percentage cropping attack)

Recovery document image		PSNR
25%	 <p>Recently, wireless mesh networks is attracting significant interest from academia, industry, and standard organizations.</p>	17.712
50%	 <p>Recently, wireless mesh networks is attracting significant interest from academia, industry, and standard organizations.</p>	14.655

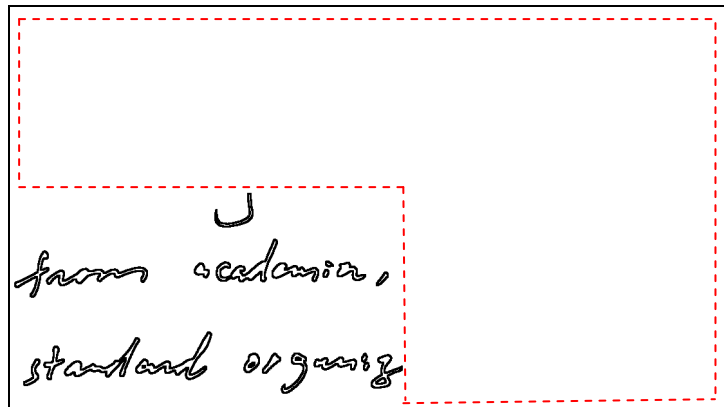
75%		13.585
-----	--	--------

## 4.7. Embedding Capacity

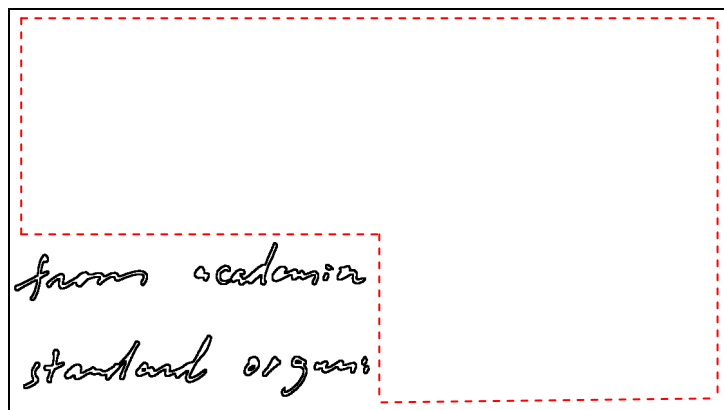
In this section, we conduct experiments to compute the embedding capacity for the proposed method. We used handwritten document images as the experimental objects. The total information size of black skeleton is 15168 bits as in Fig. 11(b). Fig. 13 indicates all of the embeddable locations for the handwritten document. There are totally 60250 bits can be embedded. When stego document image was cropped out 75%, the embedding capacity remains only has 18144 bits (see Fig.14). We found that, as in Fig. 15, if the embeddable capacity (12722 bits) is less than the size of the black skeleton (15168 bits), we will not be able to extract correctly the black skeleton (see Fig. 16).



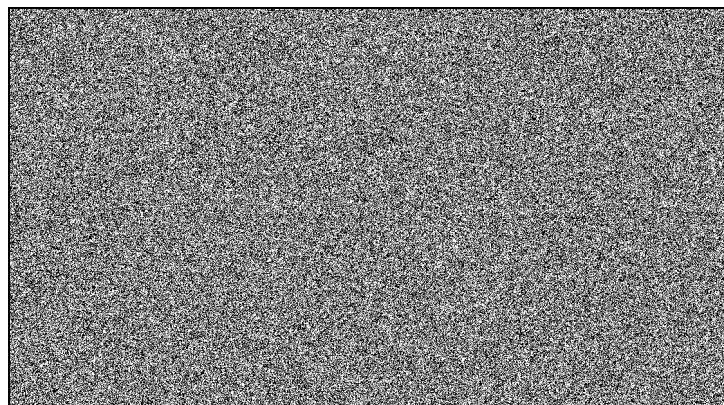
**Figure 13.** Embedding capacity (60250 bits)



**Figure 14.** Embedding capacity be cropped 75% (18144 bits)



**Figure 15.** Embedding capacity less than the size of black skeleton (12722 bits)



**Figure 16.** Incorrect black skeleton

# Chapter 5

## Conclusions and Future Works

In this thesis, we propose a novel method which can detect tampering of binary document images. By extracting the black skeletons and using information hiding technique, the binary document can be effectively recovered. The experimental results show that our proposed scheme is promising for protection of binary document images.

However, since most text documents are in binary format, thus there is not enough capacity to embed hidden information in the document unnoticeably. When we embed information into the host image, the original host image would be changed somehow. If the information hiding process does not effect the original information of the document image, then our proposed method is experimentally found to be very effective in protecting and recovering of binary document images from tampering.

At present, visual quality of recovered stego images are not very good. In the future, we hope to improve the quality of the black skeleton of the skeleton extraction process, and so to enhance the visual quality of recovered images.

Also, in this thesis, we focus on robustness of our scheme against the cropping attack. We can combine the white skeleton with the black skeleton to have a more robust scheme for binary document image against other attacks.

# Bibliography

- [1] C. H. Lin, W. K. Chang, Y. Y. Lin and L. Y. Cheng, "Binary Document Images Authentication by Thinning Digital Patterns," The Third International Conference on Intelligent Information Hiding and Multimedia Signal, Kaohsiung, Taiwan, Nov. 2007.
- [2] Y.-F. Hsu and S.-F. Chang, "Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency," IEEE International Conference on Multimedia and Expo, July 2006.
- [3] H. Lu, A. C. Kot and J. Cheng, "Secure Data Hiding in Binary Document Images for Authentication," Proceedings of the 2003 International Symposium on Circuits and Systems, Vol. 3, May 2003, Bangkok, Thailand, pp:806-809.
- [4] H. Yang and A. C. Kot, "Data Hiding for Text Document Image Authentication by Connectivity- Preserving," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, March 18-23, 2005, Philadelphia, PA, USA, pp:505-508.
- [5] Q. Mei, E. K. Wong and N. Memon, "Data hiding in binary text document," Proceedings SPIE, 2001, Vol. 4314, pp:369-375.
- [6] H. Yang and A. C. Kot, "Pattern-Based Data Hiding for Binary Image Authentication by Connectivity-Preserving," IEEE Transactions on Multimedia, Vol. 9, No. 3, April 2007, pp:475-486.
- [7] M. Wu and B. Liu, "Data Hiding in Binary Image for Authentication and Annotation," IEEE Transactions on Multimedia, Vol.6, No 4, Aug. 2004, pp:528-538.
- [8] B. Zhao, K. Fan, Y. Wang and W. Kou, "Adaptive Encrypted Information Hiding

Scheme for Image,” International Conference on Computational Intelligence and Security, Dec. 2007, pp:950-953

- [9] C. H. Lin, J. S. Chou and Y. W. Chen, “Integrity Protection of Document Assets by Computing Skeletons,” The Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), Kumamoto, Japan, Sept 2007.
- [10] Neil F. Johnson, Sushil Jajodia, “Exploring Steganography : Seeing the Unseen,” IEEE Computer, Vol. 31, No. 2, Feb. 1998, pp.26-34.
- [11] B. Pfitzmann, “Information Hiding Terminology, ” Proceedings of the first workshop on Information Hiding,” Lecture Notes in Computer Science, Springer-Verlag, Berlin, Cambridge, UK, pp.347-350.
- [12] F. A. P. Petitcolas, R J. Anderson, M. G. Kuhn, “Information Hiding –A Survey,” Proceedings of the IEEE, Vol.87, No.7, July 1999, pp.1062-1078.